2013

Prophage Proximities To Cell Boundary Genes And Other Functional Categories In Escherichia Coli Genomes

Corey D. Young

North Carolina Agricultural and Technical State University

Prophage Proximities to Cell Boundary Genes and Other Functional Categories in *Escherichia coli* Genomes

Corey D. Young

North Carolina A&T State University

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department: Biology

Major: Biology

Major Professor: Dr. Scott H. Harrison

Greensboro, North Carolina

2013

The Graduate School
North Carolina Agricultural and Technical State University
This is to certify that the Master's Thesis of

Corey D. Young

has met the thesis requirements of
North Carolina Agricultural and Technical State University

Greensboro, North Carolina
2013

Approved by:

Scott Harrison, PhD Major Professor	Dukka KC, PhD Committee Member
Gregory Goins, PhD Committee Member	Perpetua Muganda, PhD Committee Member
Mary Smith, PhD Department Chair	Sanjiv Sarin, PhD Dean, The Graduate School

Biographical Sketch

Corey D. Young is a native of Roanoke Rapids, NC. He was born August 18, 1988 into a loving family who supports him in all of his endeavors. As the oldest of two, Corey attended Halifax County Schools through high school, where he was a standout athlete. As a sophomore in high school, a young science teacher sparked Corey's interest. Corey's parents, Curtis and Faye Young, pushed Corey to register for courses that peaked his interest, which included AP Biology, Physics, and Calculus. His younger brother, Chase P. Young, is currently a senior in the speech pathology department at North Carolina A&T State University. Corey seeks a visual and tangible proof of his contributions to society, as evidence by the quality of life experienced by people he encounters. Corey's strengths include his commitment, aptitude necessary to achieve his goals, patience, and a determination to be successful. Corey is quiet and unassuming by nature. His choice to embark and complete undergraduate and graduate degrees in biology at North Carolina A&T State University prove his growth not only as a scholar but also as an individual and citizen. Working closely under the direction of Dr. Scott H. Harrison, Corey has sharpened both his computational aptness and overall understanding of biology. Given the opportunity, Corey is interested in pursuing career goals that address complex challenges at the intersection of science and industry.

Dedication

This thesis is dedicated to my family, friends and support team here at North Carolina A&T State University. You have all encouraged me to achieve goals that I had not even set for myself. Thank You.

Acknowledgements

I would like express the upmost gratitude and appreciation to my major professor, Scott H. Harrison, PhD. Without his attitude, advice and instruction, this thesis could not be possible. I would like to thank my committee members, Gregory Goins, PhD, Dukka KC, PhD, and Perpetua Muganda, PhD, whose critiques not only bettered my work but also increased my overall understanding of the topic at hand by supporting my engagement of the scientific literature.

This material is based in part upon work supported by the National Science Foundation under Cooperative Agreement No. DBI-0939454. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

List of Figures viii

List of Tables ix

Abstract 2

CHAPTER 1 Introduction 3

1.1 Hypothesis and Rationale 3

CHAPTER 2 Literature Review 5

2.1 Temperate Phages 5

2.2 Phage Effects on Bacterial Virulence 5

2.3 Bacterial Invasion 6

2.4 Cargo Genes and Genomic Re-organizations due to Phages 7

2.5 Categorization of Gene Function 8

2.6 Detection of Phage Insertions 9

2.7 Reproducibility and Genomic Data Workflows 10

2.8 Potentials and Challenges of Phylogenetic Analyses 11

CHAPTER 3 Materials and Methods 13

3.1 Materials 13

3.2 Data Curation 14

3.3 Data Analysis 16

3.4 Workflow Implementation 17

CHAPTER 4 Results 18

4.1 Data Integration and Processing 18

4.2 Frequency of Genomes with Intact Prophage Counts 18

4.3 Phylogenetic Analysis of 49 Genomes and Controlled Comparison of 10 Genomes 20

4.4 Odds Ratios for Genomic Positions of Functional Gene Categories 21

4.5 GALAXY Workflow 24

CHAPTER 5 Discussion and Conclusion 27

References 30

Appendix 38

List of Figures

Figure 1. Schematic for data integration and processing. 18

Figure 2. Frequency distributions of Escherichia coli strains based on number of prophages and pathogenicity status. 19

Figure 3. Molecular phylogenetic analysis by Maximum Likelihood method. 20

Figure 4. Odds ratios of different COG-categorized sets of genes in 49 genomes per proximity to intact prophages. 22

Figure 5. Odds ratios of different COG-categorized sets of genes in 10 genomes. 23

Figure 6. Odds ratios of different COG-categorized sets of genes in 49 genomes per proximity to quasi-prophages. 24

Figure 7. Interface for tool implemented in GALAXY: Analysis of Prophage Proximity to Gene Categories (APPGC). 25

Figure 8. Example output from the APPGC tool in GALAXY. 26

List of Tables

Table 1 Software Components 13

Table 2 Non-pathogenic Strains of Escherichia coli 14

Table 3 Pathogenic Strains of Escherichia coli 15

Table A1 Descriptions of Functional Categories for Clusters of Orthologous Groups 38

Table A2 Percentage of Genes with “Phage” in Product Name for Controlled Comparison of 10 Strains 39

Abstract

Changes in bacterial genomes due to the integration of prophages have been proposed to be part of a symbiotic relationship. Prophage activity is common for bacterial pathogens in the fluctuating environment of animal hosts. Pathogenic bacteria have been found to have extensive variation of their bacterial cell boundary zone of components, which is the inherent interface for virulence properties of antigenicity, toxicity, and resistance. Prophages are a source of this variation, both through insertion of cargo genes via prophages into bacterial strains, and additional effects due to prophage insertions affecting the overall genomic structure. The latter scenario of genomic reorganization effects has not been well explored. Our hypothesis on genomic reorganization effects was that prophage integration sites would adaptively associate with locations of cell boundary genes occurring outside the prophage insertion sites. Using clusters of orthologous groups (COG) designations, we investigated how prophage insertions collocate with COG-based categories of genes into chromosomes of pathogenic versus non-pathogenic bacteria. Here we study the integration of prophages into the genomes of 49 strains of *Escherichia coli*. The frequency of genomes containing intact prophages was much higher for pathogenic strains than for non-pathogenic strains. We examined likelihoods of proximity at which prophages integrated near genes of different COG-based categories and found that significant integration occurs near cell boundary genes. This workflow was then implemented as a tool inside the web-based genome analysis system GALAXY to enable further study of other bacterial varieties and overall genomic context.

CHAPTER 1

Introduction

There have been numerous advances in understanding the effects of viruses that infect bacteria. These bacterial viruses are also referred to as bacteriophages or phages. Virulent phages have a post-infection lytic growth mode, in which the phages directly undergo intracellular replication and lyse the cell. Temperate phages have an additional non-lytic growth mode where a post-infection stage of lysogeny takes place by which the phage replicates within the bacterial genome as a prophage (Frost, Leplae, Summers, & Toussaint, 2005). Questions remain concerning the effects of how phages alter chromosomal content by insertion of their genomic material into the bacterial chromosome (Touchon et al., 2009). This is especially important considering how prophages have been found to have a role in pathogenicity for bacteria and eukaryotic hosts (Bobay, Rocha, & Touchon, 2013; Brüssow, Canchaya, & Hardt, 2004; Wagner & Waldor, 2002). A general challenge has been to navigate the complex diversity of prophage elements that are to be found on a wide variety of bacterial genomes (Akhter, Aziz, & Edwards, 2012). There has been a corresponding development of bioinformatics tools that investigate prophage locations across bacterial genomes (Bose & Barber, 2006; Lima-Mendez, Van Helden, Toussaint, & Leplae, 2008; Zhou, Liang, Lynch, Dennis, & Wishart, 2011). As prophage data depositories expand and prophage detection tools evolve, there have been overall improvements for sensitivity, positive prediction, speed and interoperability. These are expected to enable a greater potential for integrative comparisons.

1.1 Hypothesis and Rationale

We hypothesized that the three-way association of how phages alter bacteria to affect eukaryotic hosts would include differential repositioning effects of prophage insertions near

functional categories of genes. Specifically, prophage insertions may be expected to modulate expression for functional category genes associated with pathogenesis, such as genes related to the cell boundary that, when altered, may enable the pathogen to evade host defenses. It is known that the bacterial cell boundary, when altered in terms of gene-level mutations, helps evade host defenses (Van Der Woude & Bäumler, 2004). There has not yet been a thorough analysis and automated workflow by which to address this novel proposal for phage-induced genomic reorganization achieving some similar evasive effect. The establishment of a workflow infrastructure is ultimately expected to enable broad-ranging analysis across diverse bacterial species and positional effects of prophages that relate to emerging knowledge of regulatory outcomes of genomic reorganization. Our approach was to then implement an overall software architecture that would be extensible for a scalable analysis. Based on this objective, a tool was constructed for operation within the web-based platform GALAXY, allowing for a data-intensive genomic analysis (Blankenberg et al., 2010; Giardine et al., 2005; Goecks, Nekrutenko, Taylor, & Team, 2010).

CHAPTER 2

Literature Review

2.1 Temperate Phages

Temperate phages have variable genomes, where phage homologies across a group of closely related bacterial host strains are less than virulent phage homologies where the homology ranges are <60% and >80% respectively. This difference appears to occur across bacterial diversity, having been found for *Bacillus subtilis*, *Escherichia coli*, *Lactococcus lactis*, *Streptococcus thermophilus* and species of *Mycobacterium* (Chopin, Bolotin, Sorokin, Ehrlich, & Chopin, 2001). Temperate phage infection of a bacterium occurs through two pathways once in the cell: the lytic pathway in which lysis occurs because of a vast production of viral particles, and the lysogenic pathway in which the cell survives with the lytic capacity of the virus turned off (Echols, 1972). An advantage of the lysogenic cycle is that it allows for persistence of the virus without exhausting the supply of bacterial host cells which would otherwise result from an unchecked series of lytic infections (Echols, 1972). There are two primary events associated with lysogeny: repression of genes for lytic functions, and integration of the viral DNA in the host DNA (Echols, 1972). A well-studied temperate phage is lambda phage where both production of the cl repressor and integrase are found only for the lysogenic cycle and not the lytic cycle (Maloy & Freifelder, 1994). A lysogenic mode of infection continues until the expression of lytic cycle genes, including those that would introduce site-specific recombination events for the excision of the prophage.

2.2 Phage Effects on Bacterial Virulence

For the virulence of bacteria, a famous claim in the history of microbiology has been that "…the actions and reactions are not solely between these two beings, man and bacterium, for the

bacteriophage also intervenes" (d'Herelle, 1930; Wagner & Waldor, 2002). Early investigators exposed nontoxigenic streptococci to toxigenic streptococcal and found that nontoxigenic cultures acquired the ability to produce scarlatinal toxin (d'Herelle, 1930). Experimental work demonstrated that bacteria had a filterable agent that transmitted virulence properties (Frobisher & Brown, 1927). This ability for acquiring virulence was later found for multiple varieties of bacteria, and was later attributed to phages in a phenomenon now known as transduction. Heat shocked supernatants of filtered cellular suspensions were found to contain phages that were transferring genetic material from one cell to another (Zinder & Lederberg, 1952). A wide range of genes that encode virulence properties have been found to undergo transfer by transduction in bacteria (Wagner & Waldor, 2002). Toxin genes are a common type of virulence factor that may be encoded as "cargo" by bacteriophage, but other examples of identified virulence factors have been regulatory factors that increase virulent gene expression and a range of structural components for successful colonization of animal host (Wagner & Waldor, 2002). In terms of the different virulence factors for adhesion, colonization and invasion, the cell boundary is an essential aspect to how bacteria interface with multicellular organisms, and may therefore be considered a promising area for studying the overall phenomenon of phage-induced bacterial virulence.

2.3 Bacterial Invasion

Microbial pathogens have evolved a variety of ways to invade the host and survive, avoid and/or resist immune response, damage cells, and multiply in specific and normally sterile regions (Cossart & Sansonetti, 2004). Some bacteria have been found to induce their own uptake into the nutrient-rich intracellular environment of eukaryotic cells (Cossart & Sansonetti, 2004). Invasion into animal host cells typically requires interaction between bacterial surface protein

adhesins and animal cell surface receptors (Kuespert, Weibel, & Hauck, 2007). To gain entry into a host cell, many invasive bacteria exploit the molecules of cellular adhesion as much as they exploit the host cell machinery (Kuespert et al., 2007). Experimental investigations have included transposon mutagenesis where knockout mutations in *E. coli* have revealed proteins like OmpA to be necessary for the invasion of endothelial cells in vitro and in vivo (Huang et al., 1999). The invasiveness of knockout mutants were significantly diminished when compared to the parental strain for invading brain microvascular endothelial cells in vitro and infection of the central nervous system in vivo (Huang et al., 1999), and this associated with an overall reduced occurrence of meningitis. These phenotypic outcomes, combined with a cataloguing of other transduced genes which encode a range of other products found on the cell boundary (Brüssow et al., 2004), suggest that the alteration of cellular composition by transduction to affect host cell interaction is a major mechanism for bacterial pathogenesis.

2.4 Cargo Genes and Genomic Re-organizations due to Phages

Phages are a major cause of genetic variation for bacterial populations (Thomson et al., 2004). Bacterial lineages alter genetic material in two primary ways: slowly through vertically inherited mutations, or quickly through horizontal transfer. The rapid evolution driven by horizontally transferred genes provides an advantage for bacterial lineages in rapidly fluctuating environments (Brüssow et al., 2004). The increased rate of horizontal gene transfer due to prophage insertions is attributed to those transmitted "cargo" genes that encode traits adaptive to the host, many of which are virulence factors in bacterial pathogens (Bobay et al., 2013; Brüssow et al., 2004). A common assumption is that either improper prophage excision or illegitimate recombination results in the inaugural formation of cargo genes, but their effects on bacterial adaptation seem to be the key to the continuation of their presence and potential for further

horizontal transfer within the bacterial population (Perkins et al., 2009; Tóth et al., 2009). A generally accepted model is that the adaptive selection of cargo genes increases phage fitness indirectly based on increased fitness for the bacterial host (Desiere, McShan, van Sinderen, Ferretti, & Brüssow, 2001; Hendrix, Lawrence, Hatfull, & Casjens, 2000). Variation across phages has been found to be intensified due to the shuffling of phage modules and cargo genes in different *E. coli* (Brüssow et al., 2004; Mead & Griffin, 1998). Although the symbiosis that has been proposed for prophages and their bacterial hosts has been touted to minimize disruptions to genomic structure (Bobay et al., 2013; Brüssow et al., 2004), prophage disruptions in the bacterial genome may also affect organizational traits such as genes encoding functional neighbors, transcriptional controls, supercoiling-related expression effects, genes congregating close to the origin of replication, and the interdependencies between many regulatory signals (Bobay et al., 2013; Brüssow et al., 2004; Couturier & Rocha, 2006; Lathe III, Snel, & Bork, 2000; Rocha & Danchin, 2003; Touzain, Petit, Schbath, & El Karoui, 2010). In summary, most research on phage effects in bacteria has generally involved cargo genes. By comparison, the investigation of re-organization outcomes due to prophage insertions is a frontier area that is ripe for further investigation.

2.5 Categorization of Gene Function

With the abundance of genomic sequences, there has been a pressing need for an exhaustive cataloguing of genetic function. The functional cataloguing of potential gene regions has been often pursued through sequence similarity inferences of function with other genes having experimentally established functions. Orthologous relationships between genes have been used to establish evolutionary origins between genes of diverse lineages that have subsequently been clustered into functional categories (Wall, Fraser, & Hirsh, 2003). One of the first systems

to connect the vertical origins of gene histories with function is the Cluster of Orthologous Groups (COGs) where "each COG contains conserved genes from at least 3 phylogenetically distant clades and accordingly, corresponds to an ancient conserved region (ACR)" (Tatusov, Koonin, & Lipman, 1997). A final version of the original NCBI-hosted COG collection consisted of 138,458 protein-coding genes from 66 genomes categorized into 487 COGs – a classification strategy that mapped general functional categories to 75% of the annotated protein-coding genes (Tatusov et al., 2003). Another approach for comparing proteins to previously identified sequences of similar proteins and grouping them into kinship-based protein families is the Pfam database (Punta et al., 2012).

The United States Department of Energy Joint Genome Institute's Integrated Microbial Genomes (IMG) database strives to designate protein annotations to each gene from classification systems such as COG, Pfam, TIGRfam and InterPro (Chen et al., 2013). There are other powerful approaches for categorizing gene function that draw from complex hierarchical and pathway-based classifications such as Gene Ontology (GO) (Ashburner et al., 2000) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012). COG categorizations have had, overall, a lengthy historical period of systematic application to well-studied organisms such as *Escherichia coli*. Although simple in comparison to GO and KEGG, the COG categorization approach remains implemented in a generally robust and uniform manner across the genomes of the IMG database, and includes a functional category (COG M) that is distinctive for genes associated with the cell boundary zone.

2.6 Detection of Phage Insertions

Hefty portions (>20%) of the bacteria genome have been attributed to functional and non-functional genes of prophage insertions, and these insertions can often account for variation in

closely related clades or species (Casjens, 2003). Both experimental and computational approaches have been used to identify prophages. The experimental approach can only affirm the presence of a viable phage but not that of a defective prophage (Casjens, 2003). In terms of computational approaches – that are especially advantageous in this era of high volume DNA sequencing – there are increasingly integrated and comprehensive approaches to prophage detection that have included identification strategies based on comparison to known bacterial genes and attachment site recognition algorithms (Zhou et al., 2011). This general method for prophage detection has been used in multiple programs such as Prophinder, Prophage Finder, Phage_Finder, and PHAST (Zhou et al., 2011). PHAST (PHAge Search Tool) has generally surpassed other software applications of prophage detection based on objectives for accuracy, speed and richness of annotation. PHAST can use raw or annotated bacterial genome sequence data, and can complete its analysis in 3 minutes instead of the 30-minute to 2-hour analysis time of other contemporary phage-detection software applications and is about 10% more accurate. Other enhanced features for PHAST include gene sequence input methods, graphical output which provides circular and linear genomic views, detailed and summary files, and scriptable operation (Zhou et al., 2011). Other tools are more limited and have a major, extra requirement put on the user for extensive, prior annotation of sequence (Zhou et al., 2011).

2.7 Reproducibility and Genomic Data Workflows

A recent study of 18 articles published in *Nature Genetics* found that over half of the data analyses selected could not be reproduced (Ioannidis et al., 2008). Barriers to reproduction included the absence of raw data, incomplete protocols for data processing, and an omission of detail on software and hardware that were utilized (Ioannidis et al., 2008). GALAXY, a web-based genome analysis tool that, in comparison to other software platforms with similar features

is both very flexible and free for usage (Goecks et al., 2010; Néron et al., 2009; Reich et al., 2006). GALAXY's generation of metadata and its grouping of analytical workflows into histories help ensure that computational analyses are reproducible (Goecks et al., 2010). Investigational power is not only due to its simple interface, but is also due to its ability to conjoin experimental analysis with statistical analysis, its ability to manage large amounts of data, and its interfacing to traditional data depositories (Blankenberg et al., 2010; Goecks et al., 2010). Overall, GALAXY is intended to provide a seamless cycle of use, from data analysis creation, annotation and reutilization. GALAXY's use of a complete web-based approach enables users to create web-accessible documents with embedded datasets, analyses, and workflows. By comparison, tools like GenePattern are based on a Microsoft Word 'plugin' feature (Goecks et al., 2010). Although other analysis platforms like Bioconductor, BioPerl and Biopython also provide a comparable range of methodologies for analysis, these platforms are not web-based and require users to have significant programming experience (Chapman & Chang, 2000; Gentleman et al., 2004; Stajich et al., 2002).

2.8 Potentials and Challenges of Phylogenetic Analyses

Many methods have been used to model the genetic diversification of bacterial strains, but there are not many methods established for modeling phenotypic variation (Selander et al., 1986). The traditional divisions of *E. coli* (A, B1, B2, D and E) that were first established by multilocus enzyme electrophoresis (MLEE) are often associated with pathogenicity and niche adaptation (Achtman et al., 1986; Leopold, Sawyer, Whittam, & Tarr, 2011; Selander et al., 1986). This approach has not been robust across different techniques. There are differences in phylogenetic topologies of multilocus sequence typing and MLEE, in which *E. coli* groups branch differently dependent on the method used (Herzer, Inouye, Inouye, & Whittam, 1990).

Single-gene phylogenies, no matter how well conserved (e.g., 16S rRNA), fail to convey a single topology across all depths of branching associated with *E. coli* diversification. Other efforts have used whole genome sequences to build phylogenies but, due to the high frequencies of recombination across *E. coli* strains, the use of the total-genomic sequence may be counterintuitive and offer less evidence in interpreting species topology (Leopold et al., 2011). Beyond the choice of data for phylogenetic reconstruction, there are have been different perspectives on the best mathematical method to use for reconstruction. The maximum likelihood method is a frequently utilized option for reconstructing phylogenies for closely related strains, but it has been subject to criticism (Felsenstein, 1981). Parsimony methods are a common alternative to the maximum likelihood method, but data with moderate to large amounts of change will typically cause this approach to fail (Felsenstein, 1978). Once constructed, a phylogenetic tree presents serious statistical challenges; species are a part of a hierarchically-structured phylogeny and cannot be analyzed as if drawn independently from the same distribution (Felsenstein, 1985). Nonetheless, both for recently diverged groups of related strains such as the B2 group of *E. coli* (Leopold et al., 2011), and for contrasts across closely related pairs of strains which are distant to other such strain pairs (Felsenstein, 1985), high levels of confidence are very attainable for scenarios of reconstruction and comparative analysis.

CHAPTER 3

Materials and Methods

3.1 Materials

A variety of software applications, data resources, and scripting languages were used to manage the data collection and analysis in a linux-based environment (Table 1).

Table 1

Software Components

Data Processing Resources	Version or Date Used	Description
Software Applications		
PHAST	Date used: 2013.09.11	Application for identifying prophage sequences in bacterial genomes
MEGA	Version: 5.2	Molecular evolutionary genetics analysis tool. Performs sequence alignments and infers phylogenetic trees
GALAXY	Version (release) 2012.09.20	An open-source web-based server bridging experimental biology and bioinformatics with innovative tools.
R Statistics	Version 2.12.1	Statistical and graphical application coding language
Data Warehouses		
IMG	Version: 4.0	The IMG system has many features for comparative investigations across bacterial genomes.
NCBI	Date used: 2013.09.19	National Center for Biotechnology Information provides access to biomedical and genomic information
Scripting Languages		
Python	Version: 2.7.3	Programming language
Perl	Version: 5.14.2	Programming language
Operating System		
Ubuntu	12.04.2 LTS	Powerful and freely accessible Linux-based operating system

3.2 Data Curation

Fully sequenced genomes were identified for 49 different *Escherichia coli* strains, based on their having a non-deprecated status and indexing in both the Joint Genome Institute's Integrated Microbial Genomes (IMG) database (Markowitz et al., 2012) and the National Center for Biotechnology Information (NCBI) (Acland et al., 2013). Metadata attributes in the IMG database were used to infer animal-host pathogenicity of these strains. The KEGG genome database and scientific literature were used to help confirm the pathogenicity status for each strain. The genomic data and metadata were organized into a core data matrix (Tables 2 and 3). Genomic data collection was mainly limited to the chromosomal content of each strain, although plasmid information was retained for future development of the workflow.

Table 2

Non-pathogenic Strains of Escherichia coli

Strain Identifier	Chromosome Accession	Pathogen Status	Intact Prophages	Quasi-Prophages
ABU 83972	NC_017631.1	Non-Pathogen	4	1
AIEC UM146	NC_017632.1	Non-Pathogen	5	2
B REL606	NC_012967.1	Non-Pathogen	2	6
BL21(DE3)	NC_012971.2	Non-Pathogen	2	4
BL21(DE3)pLysS AG'	NC_012947.1	Non-Pathogen	1	6
BW2952	NC_012759.1	Non-Pathogen	2	6
C ATCC 8739	NC_010468.1	Non-Pathogen	3	3
DH1	NC_017625.1	Non-Pathogen	1	5
K-12 substr. DH10B	NC_010473.1	Non-Pathogen	3	6
K-12 substr. MG1655	NC_000913.2	Non-Pathogen	1	9
O139:H28 E24377A	NC_009801.1	Non-Pathogen	3	6
O150:H5 SE15	NC_013654.1	Non-Pathogen	1	0
O18:K1:H7 IHE3034	NC_017628.1	Non-Pathogen	12	3
O81 ED1a	NC_011745.1	Non-Pathogen	8	4
O9 HS	NC_009800.1	Non-Pathogen	2	5
SE11	NC_011415.1	Non-Pathogen	7	1
SECEC SMS-3-5	NC_010498.1	Non-Pathogen	2	3
W, ATCC 9739	NC_017635.1	Non-Pathogen	7	3

Table 3

Pathogenic Strains of Escherichia coli

Strain Identifier	Chromosome Accession	Pathogen Status	Intact Prophages	Quasi-Prophages
55989	NC_011748.1	Pathogen	6	0
ETEC H10407	NC_017633.1	Pathogen	8	1
IAI1	NC_011741.1	Pathogen	3	0
IAI39	NC_011750.1	Pathogen	10	5
LF82	NC_011993.1	Pathogen	4	0
NA114	NC_017644.1	Pathogen	7	4
O103:H2 12009	NC_013353.1	Pathogen	11	1
O104:H4 2009EL-2050	NC_018650.1	Pathogen	6	2
O104:H4 2009EL-2071	NC_018661.1	Pathogen	8	1
O104:H4 2011C-3493	NC_018658.1	Pathogen	7	1
O111:H 11128	NC_013364.1	Pathogen	13	5
O127:H6 E2348/69	NC_011601.1	Pathogen	10	1
O157:H7 EC4115	NC_011353.1	Pathogen	14	5
O157:H7 EDL933	NC_002655.2	Pathogen	11	6
O157:H7 str. Sakai (EHEC)	NC_002695.1	Pathogen	11	5
O157:H7 TW14359	NC_013008.1	Pathogen	14	4
O17:K52:H18 UMN026	NC_011751.1	Pathogen	5	2
O26:H11 11368	NC_013361.1	Pathogen	14	5
O44:H18: 042	NC_017626.1	Pathogen	5	4
O45:K1 S88	NC_011742.1	Pathogen	7	2
O55:H7 CB9615	NC_013941.1	Pathogen	11	1
O55:H7 RM12579	NC_017656.1	Pathogen	8	1
O6:K2:H1 CFT073	NC_004431.1	Pathogen	6	2
O7:K1 CE10	NC_017646.1	Pathogen	10	5
O83:H1 NRG 857C	NC_017634.1	Pathogen	3	0
P12b	NC_017663.1	Pathogen	5	6
UMNK88	NC_017641.1	Pathogen	9	4
UTI89	NC_007946.1	Pathogen	6	1
Xuzhou21	NC_017906.1	Pathogen	10	3
O1:K1:H7	NC_008563.1	Pathogen	9	3
O6:K15:H31 536	NC_008253.1	Pathogen	1	1

For mapping phage locations in the genomes of our selected strains, the PHAST web server was used (Zhou et al., 2011). Data were collected on the PHAST prophage types of intact, questionable and incomplete. Questionable and incomplete phage counts were grouped together into a quasi-prophage category. Summary counts of intact prophages and quasi-prophages were added into the core data matrix. The PHAST data files for each genome were collected (the summary result file and the detailed file) for the purpose of collecting chromosomal coordinates of prophage locations. 16S rRNA gene sequences were downloaded from the IMG database. COG categories of functional gene annotations were identified from genome information data files downloaded from IMG. COG categories with a minimal representation in bacterial genomes (A, B, Y, and Z) were excluded from the analysis. A complete COG category list is shown in Table A1.

3.3 Data Analysis

Mann Whitney *U* tests of intact prophage and quasi-prophage frequencies across pathogenicity status were conducted, and significance was determined at $P < 0.05$. Phylogenetic analysis of 16S rRNA was conducted in MEGA version 5 (Tamura et al., 2011), with sequence alignment by MUSCLE (Edgar, 2004) and phylogenetic tree reconstruction through the Maximum Likelihood method (MLE) based on the Tamura-Nei model (Tamura & Nei, 1993). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. Trees were drawn to scale, with branch lengths measured in the number of substitutions per site. Odds ratios were calculated for 2x2 contingency matrices for each of the 19 COG categories and proximities from prophages were evaluated on a range of 500, 1000, …, 9500, 10000 base pairs.

Significance for contingency tables was evaluated by the Fisher Exact Test, followed by use of a Dunn-Bonferroni correction factor of 380 based on the 19 COG categories common to bacterial gene annotations and the 20 proximities; i.e., $P < 0.000132$. All statistical analyses were two-tailed and performed with R (version 2.12.1, The R Foundation for Statistical Computing, http://www.R-project.org).

3.4 Workflow Implementation

A centralized data repository was constructed for PHAST data, IMG gene information, and our core matrix. For implementation into GALAXY, a "tool config file" was constructed using XML. This file is used to first build the user interface and to then link the GALAXY interface to our software with analysis scripts implemented in Python, Perl and R.

CHAPTER 4

Results

4.1 Data Integration and Processing

Figure 1 describes the scope and flow of data such as overall prophage counts, prophage counts for intact and quasi-prophages relative to pathogenicity, strain counts for scatter plot generation, and other values pertinent to the entire study.

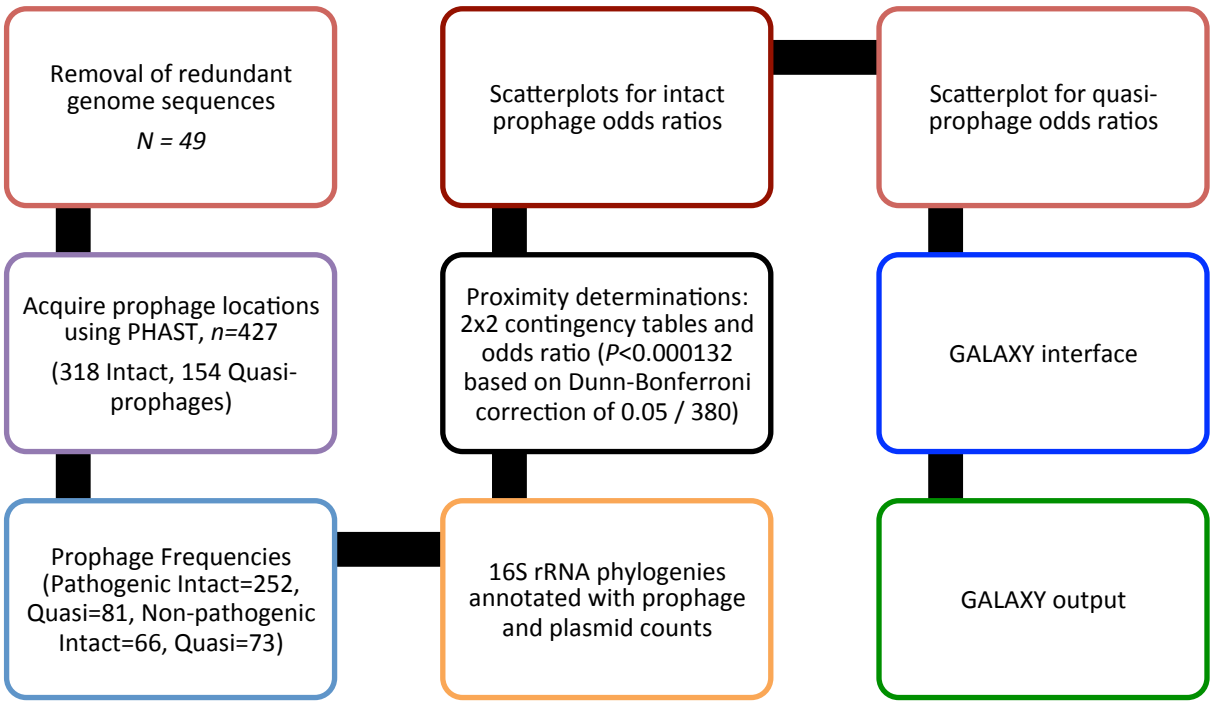

Figure 1. Schematic for data integration and processing.

4.2 Frequency of Genomes with Intact Prophage Counts

31 of our strains were classified as pathogenic and the other 18 as non-pathogenic. From all 49 strains, 472 prophages were identified: 318 intact and 154 of quasi-prophage classification. Frequencies of pathogenic and non-pathogenic strains with intact and quasi-prophages are shown in Figure 2. Comparisons on the estimated presence of prophages across chromosomal and plasmid elements of the 49 strains were conducted. Significance was found for intact prophages

(*P*<0.001, Mann-Whitney *U* test), quasi-prophage regions (*P*<0.05, Mann-Whitney *U* test), chromosome size (*P*<0.001, Mann-Whitney *U* test), and for plasmid counts (*P* < 0.05, Mann-Whitney *U* test). Intact prophage counts for genomes of pathogenic strains were 252 and 66 for those of non-pathogenic strains. Quasi-prophage counts for genomes of pathogenic strains were 81 and 73 for those of non-pathogenic strains. Figure 2 shows pathogenic strains to have a much greater number of genomes with 8 or more intact prophages compared to non-pathogenic strains.

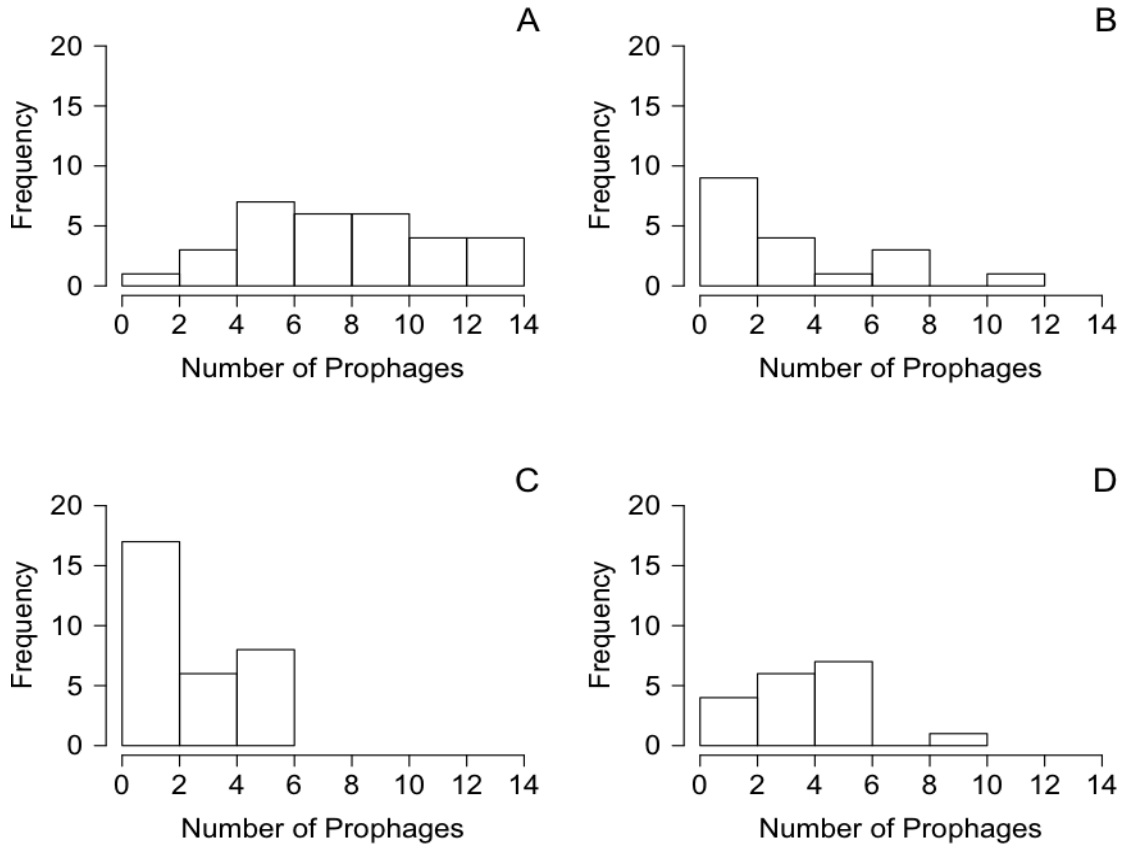

Figure 2. Frequency distributions of *Escherichia coli* strains based on number of prophages and pathogenicity status. A: Pathogenic strain subset (*n*=31) based on intact prophages. B: Non-pathogenic strain subset (*n*=18) based on intact prophages. C: Pathogenic strain subset (*n*=31) based on quasi-prophages. D: Non-pathogenic strain (*n*=18) based on quasi-prophages.

4.3 Phylogenetic Analysis of 49 Genomes and Controlled Comparison of 10 Genomes

Phylogenetic trees for the entire data set (not shown) and controlled comparison were built using MEGA5. As a diverse group of organisms, *E. coli* is categorized into four major (A, B1, B2, D) and one minor (E) group. Our choice of a single gene (16S rRNA) did not generate a phylogenetic tree representation matching that of the aforementioned *E. coli* categories. A more controlled comparison of closely related strain pairs was constructed to emulate independent comparisons. Eight paired sets having an immediate and unique last common ancestor were chosen. Five of our eight paired closely related strains had contrasting pathogenicity statuses. Of these five, only three showed an increase in prophages associated with pathogenicity (Figure 3).

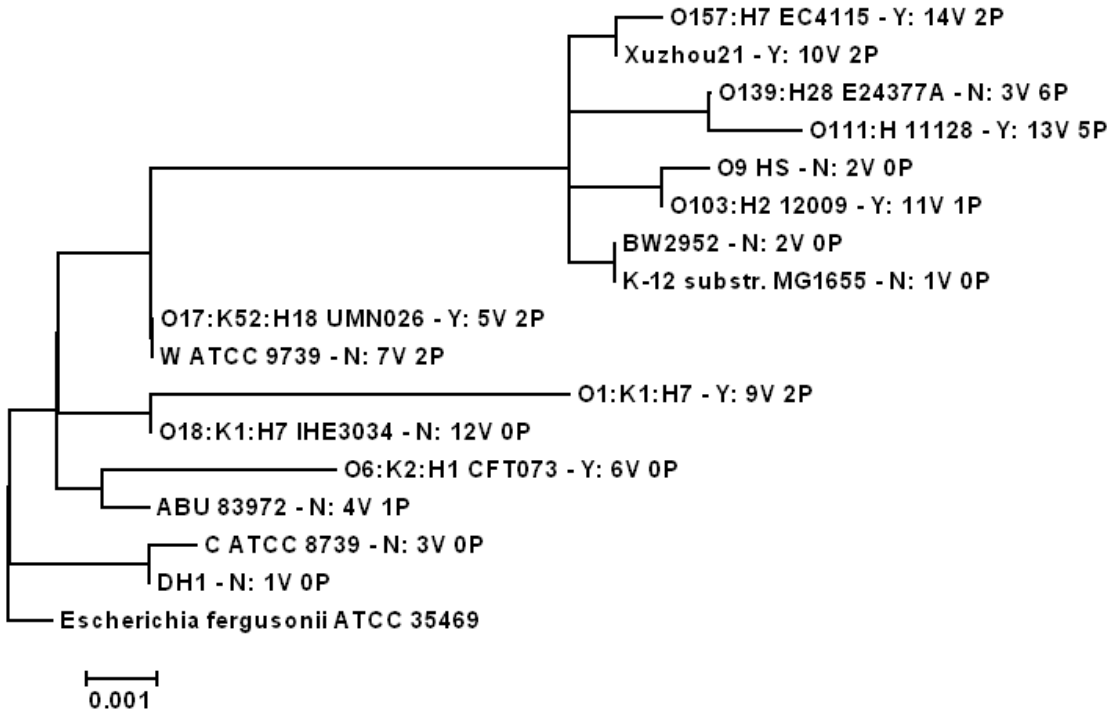

Figure 3. Molecular phylogenetic analysis by Maximum Likelihood method. For each genome, the pathogenicity status (Y-N), prophage count (V) and plasmid count (P) are shown. Evolutionary history was inferred using MLE (Tamura & Nei, 1993). The tree with the highest

log likelihood (-2459.8749) is shown. Tree is drawn to scale. The analysis involved 17 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1520 positions in the final dataset. *Escherichia fergusoni* was used to root the tree. Evolutionary analyses were conducted in MEGA5 (Tamura et al., 2011).

4.4 Odds Ratios for Genomic Positions of Functional Gene Categories

Figures 4 and 5 were constructed to show the odds ratios (O.R.) of proximities of prophage integration for functional categories based on pathogenicity status ($P<0.000132$ based on Dunn-Bonferroni correction of 0.05 / 380). As is consistent with our original hypothesis that prophage would affect the cell boundary, the COG M category (M: cell wall/membrane/envelope biogenesis) had the highest overall odds ratios. The M functional category, as it appears across 3,000 bp to 10,000 bp, suggests an especially high frequency of prophages inserted near this COG for the genomes of pathogenic strains of *E. coli* (O.R. ≈ 3). Figure 5 was constructed based on the five pairs of closely related strains from Figure 3 having a contrast in pathogenicity. Figure 5 shows that prophage insertions near COG M are 4 times more likely to significantly occur in this comparison of contrasting phylogenetic subsamples. Figure 6 portrays odds ratios for incomplete and questionable (quasi-prophage) prophages for which all are less than 1, indicating that quasi-prophages are more of a collocating factor in non-pathogenic strains for specific COG categories. Independent of prophage proximity, odds ratios for COG categories M, L and K for the genomes of the closely related strains ($n=10$) relative to pathogenicity were 0.94, 0.94 and 0.97 respectively. These odds ratios were not significant.

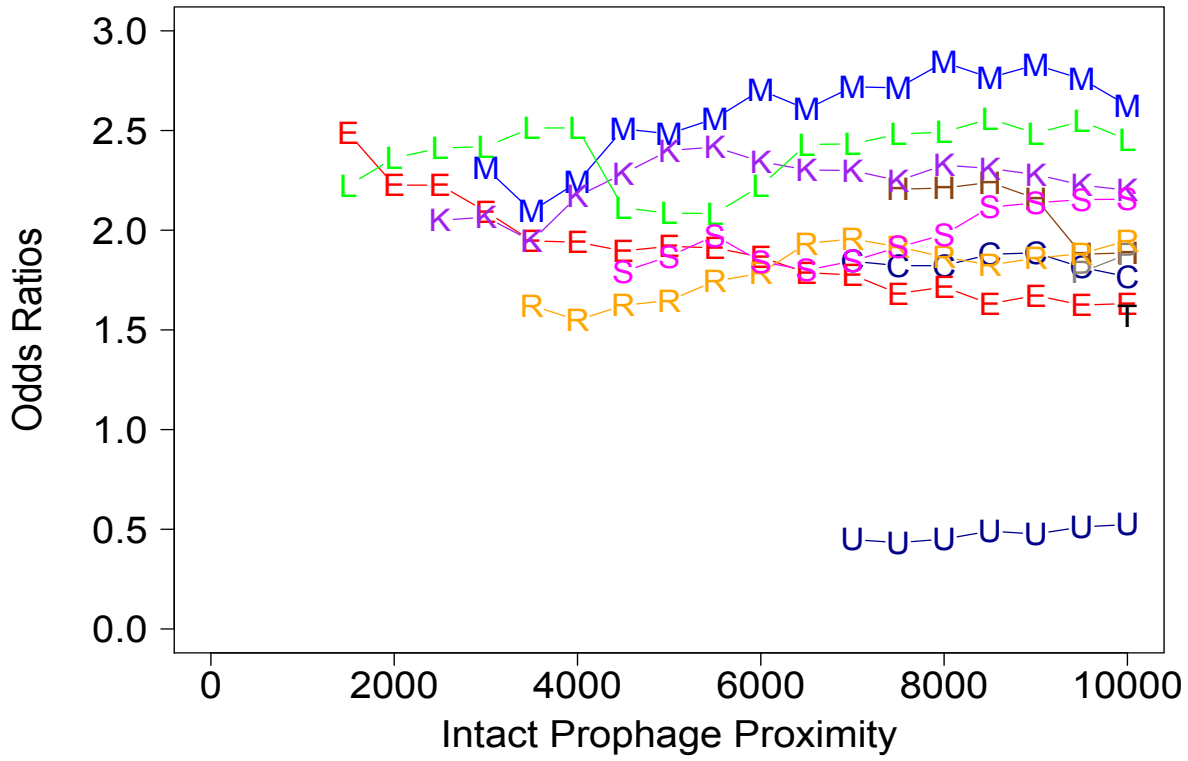

Figure 4. Odds ratios of different COG-categorized sets of genes in 49 genomes per proximity to intact prophages. Odds are for COG-categorized genes (see Table A1) being within and outside of proximities for intact prophages relative to their abundance in genomes of 31 pathogenic versus 18 non-pathogenic strains. Only significant odds ratios are shown.

A manual examination of those genes that were outside the PHAST-predicted prophage boundaries found them to be abundant in prophage-related genes. Product names of these genes in some of the notable COG categories were, for instance, putative side tail fiber protein (COG M), putative DNA-invertase from lambdoid prophage Rac (COG L), putative CI repressor of bacteriophage (COG K). This abundance generally decreased based on distance from the prophage boundary. As a basic metric, the keyword "phage" was searched for in the gene product

names listed from the IMG genomic information data files. The percentage incidences of this keyword for COG categories M, L, and K are shown in Table A2.

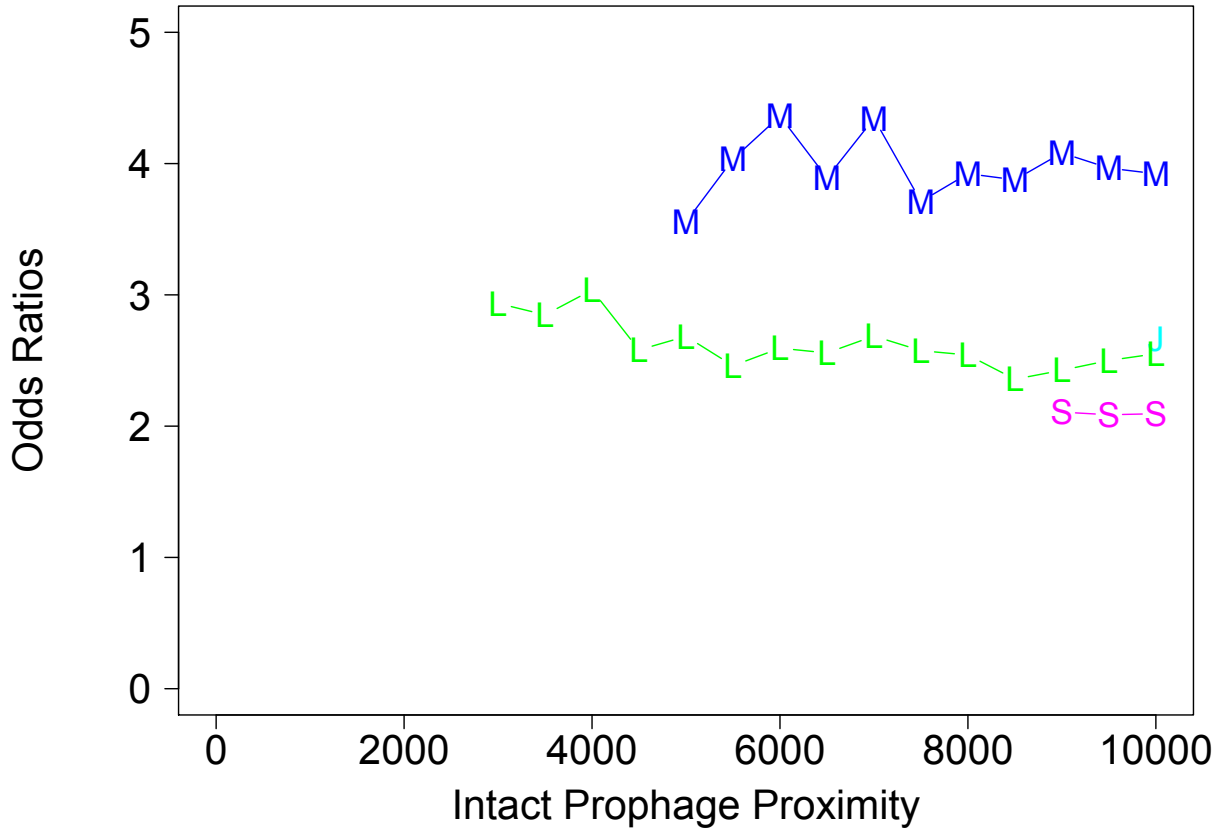

Figure 5. Odds ratios of different COG-categorized sets of genes in 10 genomes. Odds are for COG-categorized genes (see Table A1) being within and outside of proximities for intact prophages relative to their abundance in genomes of 5 pathogenic versus 5 non-pathogenic strains. Only significant odds ratios are shown.

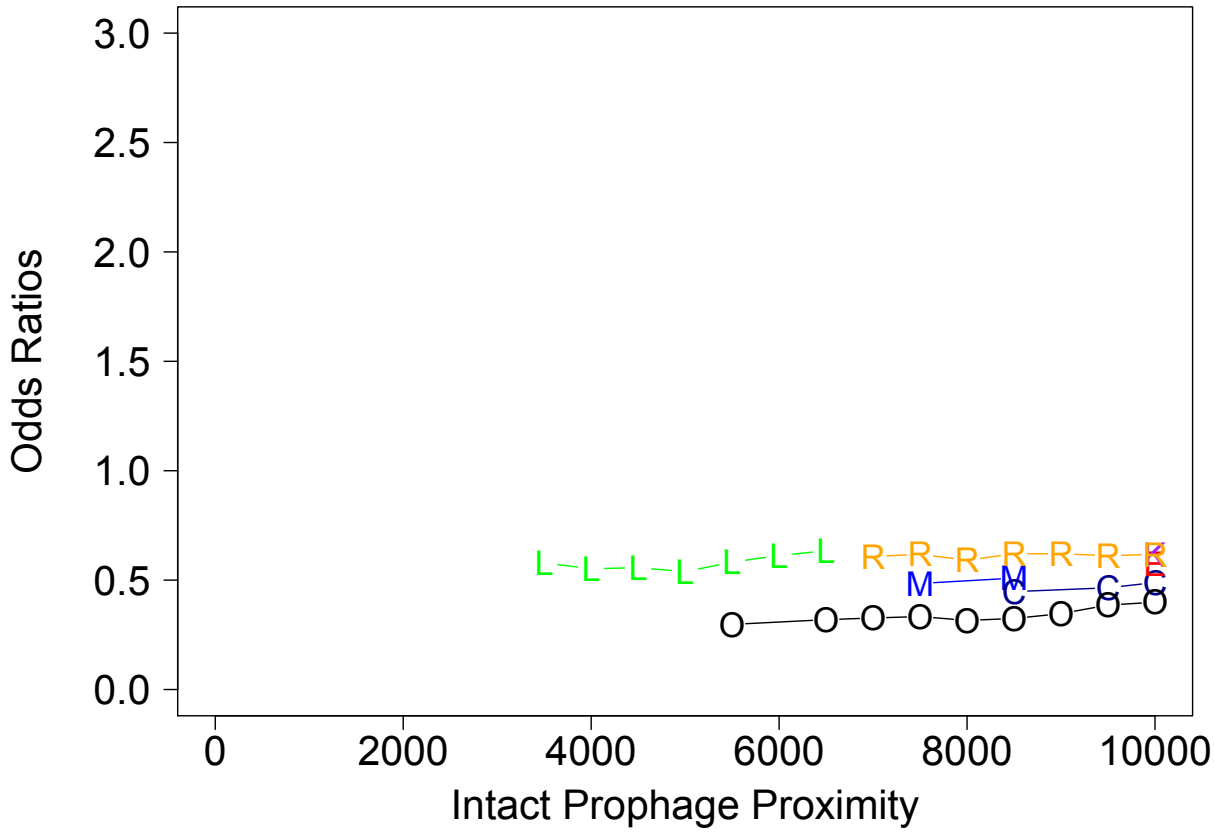

Figure 6. Odds ratios of different COG-categorized sets of genes in 49 genomes per proximity to quasi-prophages. Odds are for COG-categorized genes (see Table A1) being within and outside of proximities for quasi-prophages relative to their abundance in genomes of 31 pathogenic versus 18 non-pathogenic strains. Only significant odds ratios are shown.

4.5 GALAXY Workflow

A workflow was implemented in the web-based genomic data application, GALAXY. Our primary intention is to establish reproducibility within an extensible software framework. Figures 7 and 8 show our GALAXY interface (Figure 7) and output after an analysis (Figure 8). Users are also afforded the opportunity to select specific strain subsets, prophage type, customization of the selection of proximity values to be evaluated and modify proximity range

and proximity iterations. Prophage types are “intact prophages” and “quasi-prophages.” Modes for statistical output are a data summary mode, which enumerates the scope of the data set for pathogenic status of strains and prophage counts, and a more advanced graphical plotting of odds ratios for genomic positions of COG-categorized genes per proximity to prophage. Multiple proximities may be selected ranging stepwise from a minimal to maximal value.

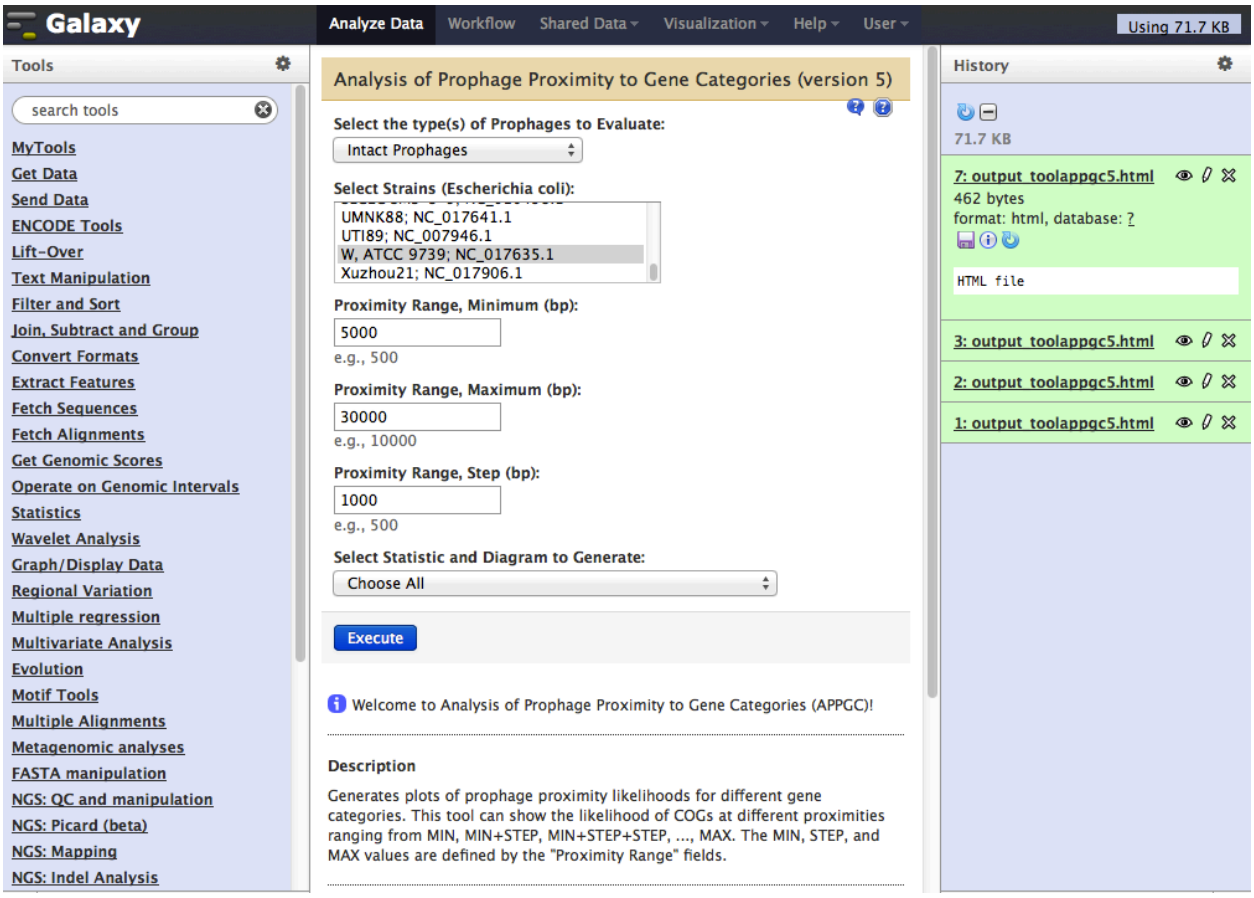

Figure 7. Interface for tool implemented in GALAXY: Analysis of Prophage Proximity to Gene Categories (APPGC). APPGC generates routine data summaries and prophage proximity plots for COG gene categories.

Figure 8 shows the output for our GALAXY tool with a prophage proximities ranging from 5,000 bp to 30,000 bp. COG M remained a dominant COG category throughout for odds ratios of proximity to intact prophages in pathogenic genomes, with an odds ratio peak greater

than 3 from 5,000 bp to 14,000 bp. Significant odds ratios of proximity to intact prophages in pathogenic genomes ranged from 1.5 to 3 for other COG categories, and also for COG M at proximities greater than 14,000 bp.

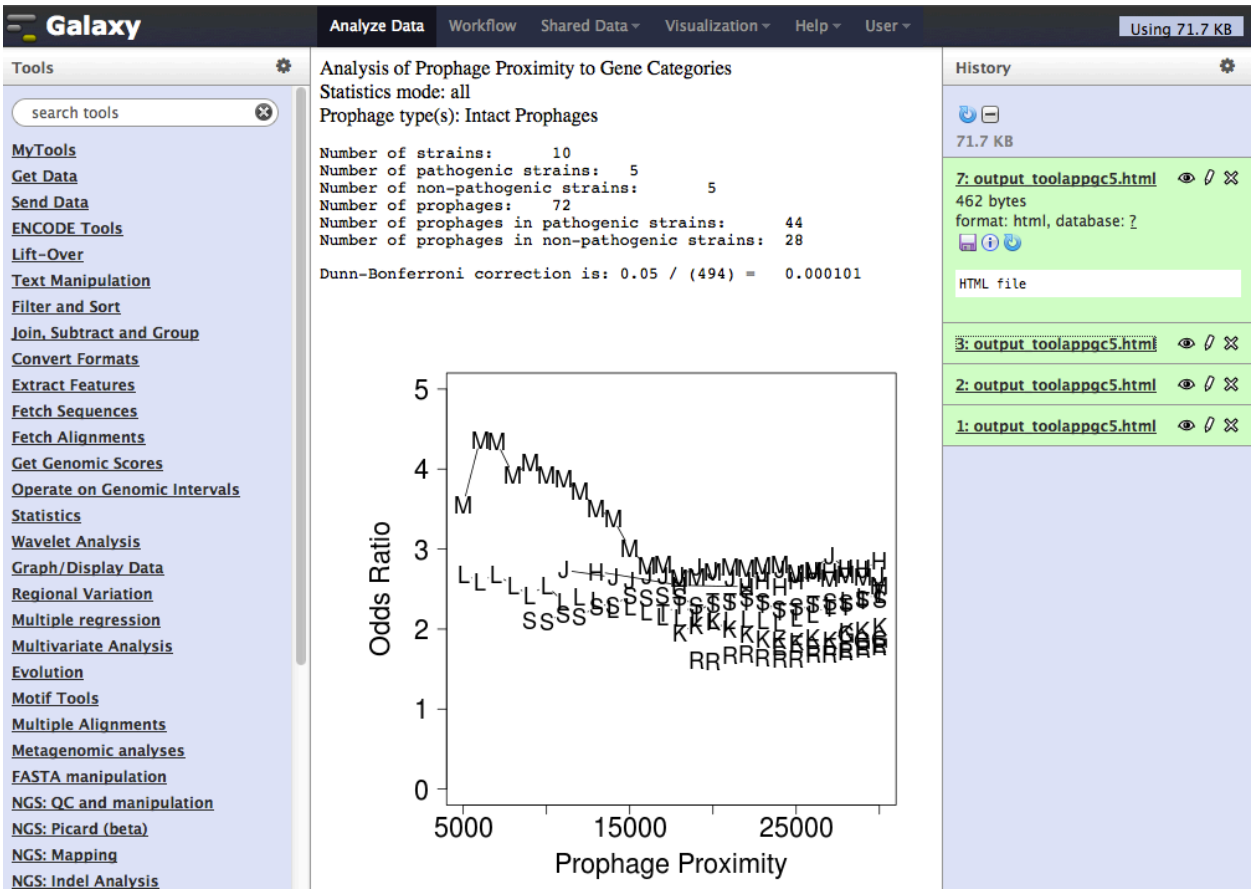

Figure 8. Example output from the APPGC tool in GALAXY. The all statistics mode was specified for analysis on the subset of strains used for the controlled comparison of 10 genomes. APPGC displays number of strains (pathogenic and non-pathogenic), number of prophages (pathogenic and non-pathogenic), Dunn-Bonferroni correction, and a graphical plot of odds ratios relative to proximities for different COG-categorized genes.

CHAPTER 5

Discussion and Conclusion

Previous studies have concluded that phages influence bacterial adhesion, colonization and invasion (Wagner & Waldor, 2002). Alterations of host bacteria that are problematic for human health include effects for resistance to immune defenses, sensitivity to antibiotics, and transmissibility among humans (Wagner & Waldor, 2002). We have developed a basic workflow for mapping phage-based alterations to the genomic organizations found for an initial set of *Escherichia coli* strains. Our implementation of this workflow in GALAXY will help expand this analysis into a wider set of strains and investigation of chromosomal context necessary for the integrative approaches required for a co-evolutionary analysis (Brüssow et al., 2004). Prophage genes can account from 10-20% of a bacterial genome and are considerable contributors to differences both within and between species (Casjens et al., 2000). Many temperate phages have been found to typically insert prophages at tRNA gene sites (A. Campbell, 2003; A. M. Campbell, 1992). We do not question this tRNA-related phenomenon, but seek to investigate further the underpinnings of phage biology by investigating proximities to other surrounding genes categorized by functionality. Our initial hypothesis was that prophage integration would affect regulatory expression of cell boundary genes through nearby insertion. The overall outcome of this study was to demonstrate a high frequency of prophages being inserted within 10,000 bp of genes for COG category M, which is the category for genes having cell wall, membrane and envelope functions. Prophage frequencies for each genome based on pathogenicity status and prophage status were examined in this study (Figure 2) and the general outcome found, for intact prophages being prominent in the genomes of pathogenic strains, was generally expected (Boyd, Davis, & Hochhut, 2001; Cheetham & Katz, 1995; Wagner &

Waldor, 2002). A phylogenetic tree analysis of our *Escherichia coli* strains challenged our initial assumption that a simple prophage count would consistently indicate a "switch" between closely related strains from a non-pathogen to that of a pathogen or vice versa (Figure 3). This issue may however be revisited with larger data sets in the future along with sophisticated approaches to matching patterns of prophage insertions to nested differences in a phylogenetic analysis (Felsenstein, 1985). There have for instance been a variety of challenges with taxonomic issues of phage classification, and problems arising from including incomplete prophages. Previous studies have indicated that some seemingly incomplete prophages may have integration/excision systems (A. Campbell, 2003). Several elements of the *E. coli* genome appear to be phage-derived but are not similar enough to prophages to be classified (A. Campbell, 2003). The rate of bacteria-prophage co-evolution accompanied with mutation, recombination and lack of universal genes have been proposed to render classical phylogenetic procedures of little use (Bobay et al., 2013). Our study demonstrated however that the simple and elegant strategy for emulating independent selections of closely related strain pairs from an overall phylogenetic tree (Felsenstein, 1985) may help uncover co-evolutionary trends. There are different phenomena that remain to be examined for subcategories within COG M, which may include genes responsible for antigenic variation, eukaryotic host cell attachment, and export of toxins. The outcomes for exact prophage insertion and potential modulations of regulatory expression for these different subcategories may be further revealed through both computational and laboratory-based approaches. The prevalence of prophage-related genes outside of predicted prophage regions (Table A2) may be due to the possibility of incorrect prophage boundary predictions or some dynamic of recombinative and selective mechanisms leading to the reinsertion of prophage genes in regions outside, but near to, intact prophages. Further analysis to distinguish vertical

inheritance versus cargo-based origins of genes and selective pressures will require examining times of divergence, strain pathogenesis, and inspection of genomic context. In summary, our finding of higher frequencies of intact prophages in the genomes of bacteria classified as pathogenic versus those of a non-pathogenic status drove us to focus on intact prophage data as a collocating factor for pathogenicity. Interestingly however, quasi-prophage collocation was an indicator of non-pathogenicity. If these collocating indicators are eventually found to be robust across other varieties of bacteria, these trends would be useful for inferring pathogenicity as a function of genomic data. For additional expansion of this work, other factors to account for the presence of prophages in genomic data include plasmid data, a more detailed examination of proximal genes, and further investigation of PHAST prophage location predictions through the use of alternative algorithms such as Prophinder, Phage Finder and Prophage Finder (Bose & Barber, 2006; Fouts, 2006; Lima-Mendez et al., 2008; Zhou et al., 2011). Although phages have been found to minimize effects on chromosome organization by the way they integrate (Bobay et al., 2013), our results indicated a broader effect for how prophage integrations impact distinct functional categories of genes, especially for the COG M category. As data and software were implemented into a GALAXY tool, we envision that this will lead into a workflow having an immense potential for expansion. Further study may especially seek to uncover the specific alterations and selective pressures associated with the repositioning effects of prophage insertions.

References

Achtman, M., Heuzenroeder, M., Kusecek, B., Ochman, H., Caugant, D., Selander, R., . . . Orskov, F. (1986). Clonal analysis of *Escherichia coli* O2: K1 isolated from diseased humans and animals. *Infection and Immunity, 51*(1), 268-276.

Acland, A., Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., . . . Church, D. M. (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 41*(D1), D8-D20.

Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Research, 40*(16), e126-e126.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Eppig, J. T. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics, 25*(1), 25-29.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., . . . Taylor, J. (2010). Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Current Protocols in Molecular Biology*, 19.10.11-19.10.21.

Bobay, L.-M., Rocha, E. P., & Touchon, M. (2013). The Adaptation of Temperate Bacteriophages to Their Host Genomes. *Molecular Biology and Evolution, 30*(4), 737-751.

Bose, M., & Barber, R. D. (2006). Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biology, 6*(3), 223-227.

Boyd, E. F., Davis, B. M., & Hochhut, B. (2001). Bacteriophage–bacteriophage interactions in the evolution of pathogenic bacteria. *Trends in Microbiology, 9*(3), 137-144.

Brüssow, H., Canchaya, C., & Hardt, W.-D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews, 68*(3), 560-602.

Campbell, A. (2003). Prophage insertion sites. *Research in Microbiology, 154*(4), 277-282.

Campbell, A. M. (1992). Chromosomal insertion sites for phages and plasmids. *Journal of Bacteriology, 174*(23), 7495.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology, 49*(2), 277-300.

Casjens, S., Palmer, N., Van Vugt, R., Mun Huang, W., Stevenson, B., Rosa, P., . . . Dodson, R. J. (2000). A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Molecular Microbiology, 35*(3), 490-516.

Chapman, B., & Chang, J. (2000). Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter, 20*(2), 15-19.

Cheetham, B. F., & Katz, M. E. (1995). A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. *Molecular Microbiology, 18*(02), 201-208.

Chen, I.-M. A., Markowitz, V. M., Chu, K., Anderson, I., Mavromatis, K., Kyrpides, N. C., & Ivanova, N. N. (2013). Improving Microbial Genome Annotations in an Integrated Database Context. *PloS One, 8*(2), e54859.

Chopin, A., Bolotin, A., Sorokin, A., Ehrlich, S. D., & Chopin, M.-C. (2001). Analysis of six prophages in *Lactococcus lactis* IL1403: different genetic structure of temperate and virulent phage populations. *Nucleic Acids Research, 29*(3), 644-651.

Cossart, P., & Sansonetti, P. J. (2004). Bacterial invasion: the paradigms of enteroinvasive pathogens. *Science, 304*(5668), 242.

Couturier, E., & Rocha, E. P. (2006). Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Molecular Microbiology, 59*(5), 1506-1518.

d'Herelle, F. (1930). The bacteriophage and its clinical applications. *The American Journal of the Medical Sciences, 180*(4), 573.

Desiere, F., McShan, W. M., van Sinderen, D., Ferretti, J. J., & Brüssow, H. (2001). Comparative genomics reveals close genetic relationships between phages from dairy bacteria and pathogenic *Streptococci*: evolutionary implications for prophage-host interactions. *Virology, 288*(2), 325-341.

Echols, H. (1972). Developmental pathways for the temperate phage: lysis vs lysogeny. *Annual Review of Genetics, 6*(1), 157-190.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792-1797.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology, 27*(4), 401-410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution, 17*(6), 368-376.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, 1-15.

Fouts, D. E. (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research, 34*(20), 5839-5851.

Frobisher, M., & Brown, J. H. (1927). Transmissible toxicogenicity of streptococci. *Bull. Johns Hopkins Hosp, 41*, 167-173.

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology, 3*(9), 722-732.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., . . . Gentry, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology, 5*(10), R80.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., . . . Taylor, J. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Research, 15*(10), 1451-1455.

Goecks, J., Nekrutenko, A., Taylor, J., & Team, T. G. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology, 11*(8), R86.

Hendrix, R. W., Lawrence, J. G., Hatfull, G. F., & Casjens, S. (2000). The origins and ongoing evolution of viruses. *Trends in Microbiology, 8*(11), 504-508.

Herzer, P. J., Inouye, S., Inouye, M., & Whittam, T. S. (1990). Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *Journal of Bacteriology, 172*(11), 6175-6181.

Huang, S.-H., Chen, Y.-H., Fu, Q., Stins, M., Wang, Y., Wass, C., & Kim, K. S. (1999). Identification and characterization of an *Escherichia coli* invasion gene locus, *ibeB*, required for penetration of brain microvascular endothelial cells. *Infection and Immunity, 67*(5), 2103-2109.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., . . . Jurman, G. (2008). Repeatability of published microarray gene expression analyses. *Nature Genetics, 41*(2), 149-155.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research, 40*(D1), D109-D114.

Kuespert, K., Weibel, S., & Hauck, C. R. (2007). Profiling of bacterial adhesin—host receptor recognition by soluble immunoglobulin superfamily domains. *Journal of Microbiological Methods, 68*(3), 478-485.

Lathe III, W. C., Snel, B., & Bork, P. (2000). Gene context conservation of a higher order than operons. *Trends in Biochemical Sciences, 25*(10), 474-479.

Leopold, S. R., Sawyer, S. A., Whittam, T. S., & Tarr, P. I. (2011). Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*. *BMC Evolutionary Biology, 11*(1), 183.

Lima-Mendez, G., Van Helden, J., Toussaint, A., & Leplae, R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics, 24*(6), 863-865.

Maloy, S. R., & Freifelder, D. (1994). *Microbial genetics*. Jones & Bartlett Learning.

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., . . . Williams, P. (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research, 40*(D1), D115-D122.

Mead, P. S., & Griffin, P. M. (1998). *Escherichia coli* O157: H7. *The Lancet, 352*(9135), 1207-1212.

Néron, B., Ménager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., . . . Letondal, C. (2009). Mobyle: a new full web bioinformatics framework. *Bioinformatics, 25*(22), 3005-3011.

Perkins, T. T., Kingsley, R. A., Fookes, M. C., Gardner, P. P., James, K. D., Yu, L., . . . Pickard, D. J. (2009). A strand-specific RNA–Seq analysis of the transcriptome of the typhoid bacillus *Salmonella* typhi. *PLoS Genetics, 5*(7), e1000569.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., . . . Clements, J. (2012). The Pfam protein families database. *Nucleic acids research, 40*(D1), D290-D301.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., & Mesirov, J. P. (2006). GenePattern 2.0. *Nature Genetics, 38*(5), 500-501.

Rocha, E. P., & Danchin, A. (2003). Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genetics, 34*(4), 377-378.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N., & Whittam, T. S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Applied and Environmental Microbiology, 51*(5), 873.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., . . . Lapp, H. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Research, 12*(10), 1611-1618.

Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution, 10*(3), 512-526.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary

distance, and maximum parsimony methods. *Molecular Biology and Evolution, 28*(10), 2731-2739.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., . . . Nikolskaya, A. N. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics, 4*(1), 41.

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science, 278*(5338), 631-637.

Thomson, N., Baker, S., Pickard, D., Fookes, M., Anjum, M., Hamlin, N., . . . Chan, K. (2004). The Role of Prophage-like Elements in the Diversity of *Salmonella enterica* Serovars. *Journal of Molecular Biology, 339*(2), 279-300.

Tóth, I., Nougayrède, J.-P., Dobrindt, U., Ledger, T. N., Boury, M., Morabito, S., . . . Oswald, E. (2009). Cytolethal distending toxin type I and type IV genes are framed with lambdoid prophage genes in extraintestinal pathogenic *Escherichia coli*. *Infection and Immunity, 77*(1), 492-500.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., . . . Bouvet, O. (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genetics, 5*(1), e1000344.

Touzain, F., Petit, M.-A., Schbath, S., & El Karoui, M. (2010). DNA motifs that sculpt the bacterial chromosome. *Nature Reviews Microbiology, 9*(1), 15-26.

Van Der Woude, M. W., & Bäumler, A. J. (2004). Phase and antigenic variation in bacteria. *Clinical Microbiology Reviews, 17*(3), 581-611.

Wagner, & Waldor. (2002). Bacteriophage control of bacterial virulence. *Infection and Immunity, 70*(8), 3985-3993.

Wall, D., Fraser, H., & Hirsh, A. (2003). Detecting putative orthologs. *Bioinformatics, 19*(13), 1710-1711.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., & Wishart, D. S. (2011). PHAST: a fast phage search tool. *Nucleic Acids Research, 39*(suppl 2), W347-W352.

Zinder, N. D., & Lederberg, J. (1952). Genetic exchange in *Salmonella*. *Journal of Bacteriology, 64*(5), 679.

Appendix

Table A1

Descriptions of Functional Categories for Clusters of Orthologous Groups

COG	Description of Functional Category
	Processing and Information Storage
A*	RNA Processing and Modification
B*	Chromatin Structure and Dynamics
J	Translation
K	Transcription
L	Replication and Repair
	Signaling and Cellular Processes
D	Cell Cycle Control and Mitosis
M	Cell Wall/Membrane/Envelope Biogenesis
N	Cell Motility
O	Post-translation Modification, Protein Turnover Chaperones
T	Signal Transduction Mechanisms
U	Intracellular Trafficking, Secretion, and Vesicular Transport
V	Defense Mechanisms
W	Extracellular Structures
Y*	Nuclear Structure
Z*	Cytoskeleton
	Macromolecule Metabolism
C	Energy Production and Conversion
E	Amino Acid Transport and Metabolism
F	Nucleotide Transport and Metabolism
G	Carbohydrate Transport and Metabolism
H	Coenzyme Transport and Metabolism
I	Lipid Transport and Metabolism
P	Inorganic Ion Transport and Metabolism
Q	Secondary Metabolites Biosynthesis, Transport and Catabolism
	Uncategorized
R	General Function Prediction Only
S	Function Unknown

*: COGs excluded for analysis based on lack of representation in bacterial genomes.

Table A2

Percentage of Genes with "Phage" in Product Name for Controlled Comparison of 10 Strains

Distance outside of prophage boundary (bp)	COG M	COG L	COG K
5,000-9,990	6.15	6.78	9.50
10,000-14,999	3.20	7.79	4.78
15,000-19,999	2.33	4.92	6.14
20,000-24,999	3.80	2.11	15.44
25,000-29,999	2.74	0.00	4.04